# Robust distributions??

..

...

...

April 18, 2011

## Outline

- Infinite mixtures of scaled Gaussians
- Student distribution
- A more general form
- Further research

# Introduction/Background

Mixtures of scaled Gaussians

$$
\begin{aligned}
f(y) &= \int_0^\infty f(y|w)f(w)dw \\
&= \int_0^\infty N(y;\mu,\frac{\Sigma}{w})f(w)dw
\end{aligned}
$$

## Introduction/Background

TABLE 1: Scale mixture of Gaussian models.

| $z$ distribution | Multivariate scale mixture distribution |
|---|---|
| Constant (Dirac delta) $(z; \sigma^2)$ | Gaussian, $\mathbf{MG}(\mathbf{y}; \sigma^2, \boldsymbol{\mu}, \boldsymbol{\Gamma})$ |
| Exponential $(z, \lambda) = \frac{1}{\lambda}\exp\left(-\frac{z}{\lambda}\right)$ | Laplacian, $\mathbf{ML}(\mathbf{y}; \lambda, \boldsymbol{\mu}, \boldsymbol{\Gamma}) = \frac{1}{\pi^d}\frac{2}{\lambda}\frac{K_{d-1}(2\sqrt{q(\mathbf{y})/\lambda})}{\left(\sqrt{\lambda q(\mathbf{y})}\right)^{d-1}}$ |
| Gamma $(z; \alpha, \mu_z) = \left(\frac{\alpha}{\mu_z}\right)^{\alpha}\frac{z^{\alpha-1}}{\Gamma(\alpha)}\exp\left(-\frac{\alpha}{\mu_z}z\right)$ | K-distribution, $\mathbf{MK}(\mathbf{y}; \alpha, \mu_z, \boldsymbol{\mu}, \boldsymbol{\Gamma})$ $= \frac{2}{\pi^d\,\Gamma(\alpha)}\left(\frac{\alpha}{\mu_z}\right)^{(\alpha+d)/2}(q(\mathbf{y}))^{(\alpha-d)/2}K_{\alpha-d}\left(2\sqrt{\frac{\alpha\, q(\mathbf{y})}{\mu_z}}\right)$ |
| Inverse Gaussian $(z; \delta, \gamma)$ $= \frac{\delta}{\sqrt{2\pi}}e^{\delta\gamma}\, z^{-3/2}\exp\left(-\frac{1}{2}\left(\frac{\delta^2}{z}+\gamma^2 z\right)\right)$ | Normal Inverse Gaussian, $\mathbf{MNIG}(\mathbf{y}; \delta, \gamma, \boldsymbol{\mu}, \boldsymbol{\Gamma})$ $= \sqrt{2}\delta e^{\delta\gamma}\left(\frac{\gamma}{\pi\sqrt{\delta^2+2q(\mathbf{y})}}\right)^{d+(1/2)}K_{d+(1/2)}\left(\gamma\sqrt{\delta^2+2q(\mathbf{y})}\right)$ |

$q(\mathbf{y}) = (\mathbf{y}-\boldsymbol{\mu})^T\boldsymbol{\Gamma}^{-1}(\mathbf{y}-\boldsymbol{\mu})$ is the scaled squared Mahalanobis distance from the mean, with $\boldsymbol{\mu} = \mathbf{0}$ for the PolSAR case.
$K_m(x)$ is a modified Bessel function of the second kind with order $m$.

Figure: ...

## Multivariate student

$$
\begin{aligned}
t(y;\mu,\Sigma,\nu) = & \int_0^\infty \mathcal{N}(y;\mu,\frac{\Sigma}{w})\mathcal{G}(w;\frac{\nu}{2},\frac{\nu}{2})\mathrm{d}w \\
= & \frac{\Gamma(\frac{\nu+p}{2})}{\Sigma^{1/2}\Gamma(\frac{\nu}{2})(\nu\pi)^{p/2}}\left(1+\frac{(y_i-\mu)^T\Sigma^{-1}(y_i-\mu)}{\nu}\right)^{-\frac{(\nu+p)}{2}}
\end{aligned}
$$

MLE estimates:

$$
\begin{aligned}
\mu = & \sum_{i=1}^N w_i y_i / \sum_{i=1}^N w_i \\
\Sigma = & \frac{1}{N}\sum_{i=1}^N w_i(y_i-\mu)(y_i-\mu)^T
\end{aligned}
$$

where

$$
w_i = \frac{\nu+p}{\nu+\delta(y_i,\mu;\Sigma)}
$$

where $\delta(y_i,\mu;\Sigma) = (y_i-\mu)^T\Sigma^{-1}(y_i-\mu)$

## Multivariate student

- Many ways to generalise the student distribution.
- Much less interest though on alternative forms of the degree of freedom parameter

Alternative forms ...

- Standard student has one disadvantage: all its marginals are Student but have the same degree of freedom and hence the same amount of tailweight.
  - A simple example is where one variable has Cauchy tails (df=1) and another Gaussian (df=300). In this situation, "the single degrees of freedom parameter has to provide a compromise between those two tail behaviours".

## Multivariate student

Approaches for varying the degrees of freedom ...
Literature:

- One solution: take a product of independent t-distributions of varying degree of freedom but assuming no correlation between dimensions.
- Jones (2002) proposed a dependent bivariate t distribution with marginals of different degree of freedom but the tractability of the extension to the multivariate case is unclear.
- Copula??

## Multiple scaled student

We consider the parametrization of the covariance matrix based on its eigenvalue decomposition:

$$\Sigma = DAD^T$$

$$p(y_i|\omega_i, \phi) = \mathcal{N}(y_i; \mu, D\Delta_i AD^T),$$

where $\Delta_i$ is the $M \times M$ diagonal matrix whose diagonal components are the inverse weights $\{\omega_{i1}^{-1}, \dots, \omega_{iM}^{-1}\}$.
- weights are used to act on the shape of the Gaussian but not on the orientation.
Weights are in addition assumed to be independent across dimension, providing:

$$p(w; \nu) = \prod_{m=1}^{M} \mathcal{G}(w_m; \frac{\nu_m}{2}, \frac{\nu_m}{2}) \quad (1)$$

## Multiple scaled student

Using a mixture of scaled normals representation,

$$p(y; \mu, A, D, \nu_1, \ldots, \nu_M)$$

$$= \int_0^\infty \cdots \int_0^\infty \mathcal{N}(y; \mu, D\bar{\Delta}_i A D^T) p(\omega_m; \nu) dw_1 \ldots dw_M$$
$$= \prod_{m=1}^{M} \frac{\Gamma(\frac{\nu_m+1}{2})}{A_m^{1/2}\Gamma(\frac{\nu_m}{2})(\nu_m \pi)^{1/2}} \left(1 + \frac{[D^T(y-\mu)]^2_{(m)}}{A_m \nu_m}\right)^{-\frac{(\nu_m+1)}{2}}$$

$$\bar{\omega}_{im}^{(r)} = \frac{\nu_m + 1}{\nu_m + \frac{v_i^{(r)\,2}(m)}{A^{(r)}(m,m)}} \qquad (2)$$

where $v_i^{(r)}(m)$ denotes the $m^{th}$ components of vector $v_i^{(r)}$ defined by $v_i^{(r)} = D^{(r)\,T}(\mathbf{y}_i - \mu^{(r)})$ and $A^{(r)}(m,m)$ the $m^{th}$ diagonal component of $A^{(r)}$.

# Multiple scaled student: different shapes

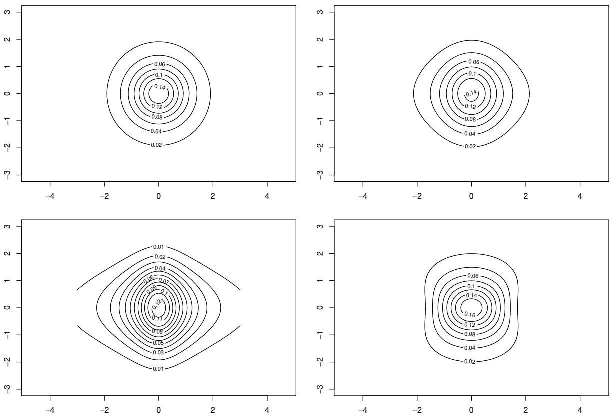

Figure: Figures of bivariate students with different degrees of freedom. Top left: v1=2, v2=2; Top right: v1=2, v2=5; Bottom left: v1=1, v2=30; Bottom right: v1=30, v2=2

# Multiple scaled student: spherical?

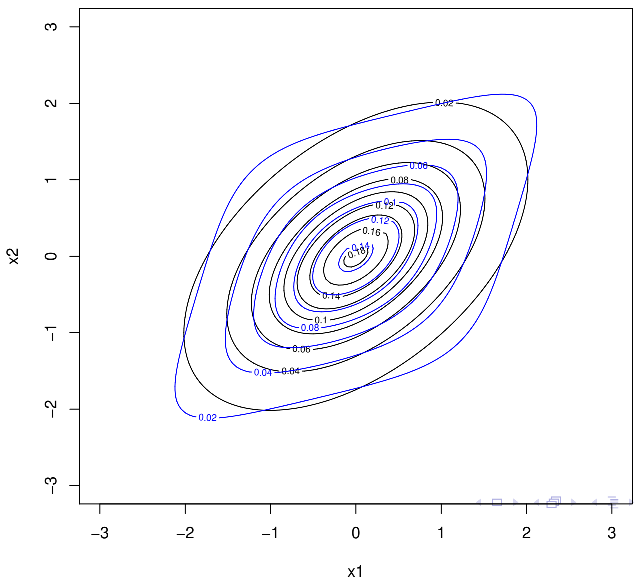


Figure: ...

# Multiple scaled student: clustering example

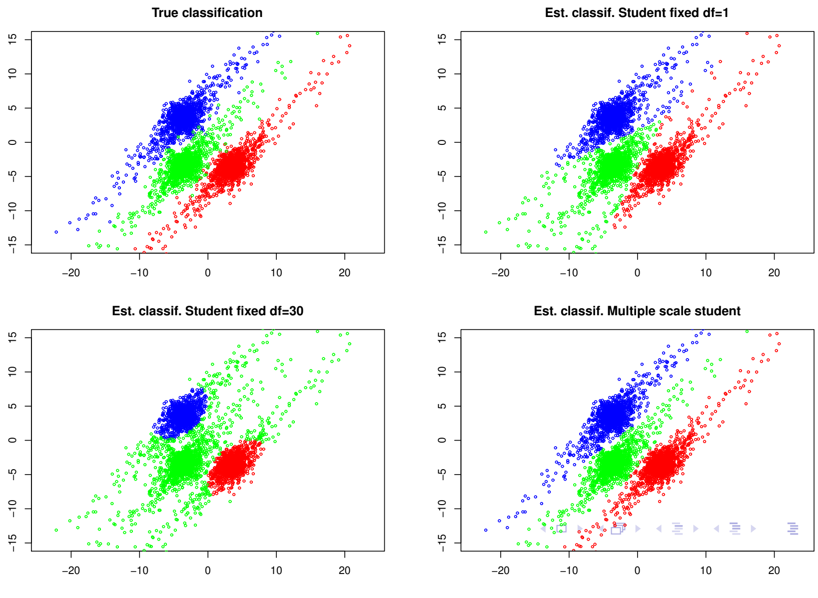


Figure: Classification results for the simulated data. Top left: True classification. Top right: Results for the student with the degree of freedom fixed to 1. Bottom left: Results for the student with the degree of freedom fixed to 30. Bottom right: Results for the multiple scaled student (df's = ...). The different colours indicate the different components to which observations are assigned to.

## Multiple scaled student

Advantages:

- Relatively straightforward estimation possible
- Medium to high dimensional data okay
- Assumes a parametric form

Disadvantages:
??

# A skewed student with multiple weights??

Hierarchical representation (scale mixture):

$$Y_i|(\gamma_{\mathbf{i}}, \omega_i) \sim N_p(\mu + \Lambda\gamma_i, \frac{\Sigma}{w_i}) \tag{3}$$

$$w_i \sim \Gamma(\tfrac{\nu}{2}, \tfrac{\nu}{2}) \tag{4}$$

$$\gamma_i|w_i \sim HN_p(0, \frac{I_p}{w_i}) \tag{5}$$

## Multiple scaled skew student

Hierarchical representation (scale mixture):

$$Y_i|(\gamma_{\mathbf{i}}, \omega_i) \sim N_p(\mu + \Lambda\gamma_i, D\Delta_i A D^T) \tag{6}$$

$$w_{im} \sim \Gamma(\tfrac{\nu_m}{2}, \tfrac{\nu_m}{2}) \tag{7}$$

$$\gamma_{im}|w_{im} \sim N(0, w_{im}^{-1})_{\mathcal{I}(0,\infty)} \tag{8}$$

## Multiple scaled skew student

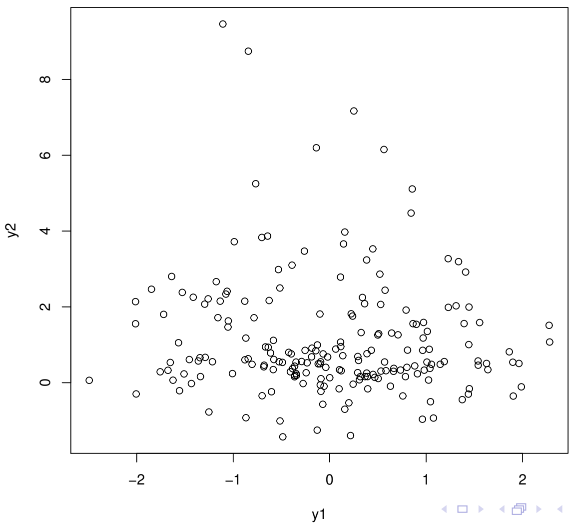


Figure: ...

# Multiple scaled skew student

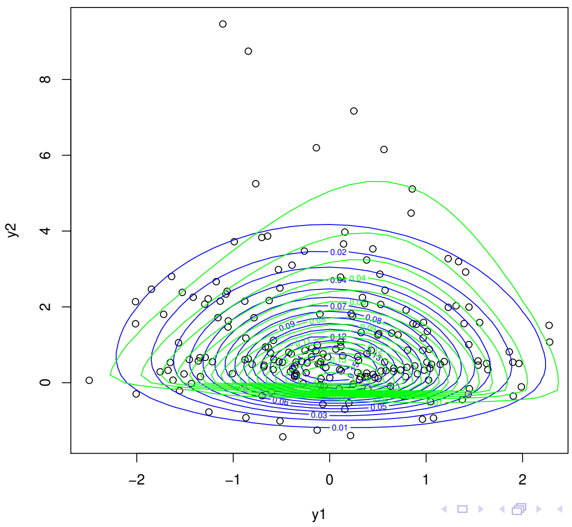


Figure: ...

## Recall: General form

Standard case:

$$f(y_i|.)) = \int_0^\infty \mathcal{N}(y_i; \mu, \frac{\sigma^2}{w_i})\mathcal{G}(w_i; \alpha, \gamma)\mathrm{d}w$$

Can $\alpha$ and $\gamma$ be different for each $i$?

$$\begin{aligned} & \int_0^\infty \mathcal{N}(y_i; \mu, \frac{\sigma^2}{w_i})\mathcal{G}(w_i; \alpha_i, \gamma_i)\mathrm{d}w \\ = \quad & \frac{\gamma_i^{\alpha_i}}{\sqrt{2}\sigma \mathrm{B}(1/2, \alpha_i)} \left[\frac{(y_i - \mu)^2}{2\sigma^2} + \gamma_i\right]^{-\alpha_i - 1/2} \end{aligned}$$

Estimation of weights,

$$w_i = \frac{\alpha_i + 1}{\gamma_i + \delta(y_i, \mu; \sigma)}$$

Priors for $\alpha_i$ and $\beta_i$?

## Further research

- Further study of the properties of the MSS
- Use of MSS for subspace clustering of high dimensional spaces?
- Application of MSS in other contexts? e.g for regression, spatial problems
- ??

## ...

Questions??